Mining the BioLiterature: towards automatic annotation of genes and proteins

Francisco M. Couto, Mário J. Silva

Departamento de Informática, Faculdade de Ciências,

Universidade de Lisboa,

Campo Grande, 1749-016 Lisboa, Portugal.

Abstract

This chapter introduces the use of Text Mining in scientific literature for biological research, with a special focus on automatic gene and protein annotation. This field became recently a major topic in Bioinformatics, motivated by the opportunity brought by tapping the BioLiterature with automatic text processing software.

The chapter describes the main approaches adopted and analyzes systems that have been developed for automatically annotating genes or proteins. To illustrate how text-mining tools fit in biological databases curation processes, the chapter presents a tool that assists protein annotation.

Besides the promising advances of Text Mining of BioLiterature, many problems need to be addressed. This chapter presents the main open problems in using text-mining tools for automatic annotation of genes and proteins, and discusses how a more efficient integration of existing domain knowledge can improve the performance of these tools.

Keywords: Text Mining, Natural Language Processing, Data Mining, Rule-based, Case-based, BioLiterature.

INTRODUCTION

Bioinformatics aims at understanding living systems using biological information. The facts discovered in biological research have been mainly published in the scientific literature (BioLiterature) since the 19th century. Extracting knowledge from such a large amount of unstructured information is a painful and hard task, even to an expert. A solution could be the creation of a database where authors would deposit all the facts published in BioLiterature in a structured form. Some generic databases, such as UniProt, collect and distribute biological information (Apweiler, 2004). However, different communities have different needs and views on specific topics, which change over the time. As a result, researchers do not look only for the facts, but also for their evidence. Before a researcher considers a fact as relevant to his work, he checks the evidence presented by the author, because facts are normally valid only in a specific context. This explains why Molecular Biology knowledge continues to be mainly published in BioLiterature. Another solution is Text Mining, which aims at automatically extracting knowledge from natural language texts (Hearst, 1999). Text-mining systems can be used to identify the following types of information: entities, such as genes, proteins and cellular components; relationships, such as protein localization or protein interactions; and events, such as experimental methods used to discover protein interactions. Bioinformatics tools to collect more information about the concepts they analyze also use Text Mining. For example, information automatically extracted from the BioLiterature can improve gene expression clustering (Blaschke, 2004).

Text Mining of BioLiterature has been studied since the last decade (Andrade and Valencia, 1998). The interest in the topic has been steadily increasing, motivated by the vast amount of publications that curators have to read in order to update biological databases, or simply to help researchers keep up with progress in a specific area. Text Mining can minimize these problems mainly because BioLiterature articles are quite often publicly available. The most widely used BioLiterature repository is MEDLINE, which provides a vast collection of abstracts and bibliographic information. For example, in 2003, about 560,000 citations have been added to MEDLINE. Reading 10 of these documents per day, it would take around 150 years to read all the documents added in 2003. Moreover, the number of new documents added per year increased by more than 20,000 from 2000 to 2003. Hence, text-mining systems could have a great impact in minimizing this effort by automatically extracting information that can be used for multiple purposes and could not possibly be organized by other means.

This chapter starts by providing broad definitions used in Text Mining and describes the main approaches. Then, it summarizes the state-of-the-art of this field and shows how text-mining systems can be used to automatically annotate genes or proteins. Next, the chapter describes a tool designed for assisting protein annotation. Finally, the chapter discusses future and emerging trends and presents concluding remarks.

TEXT MINING

Text Mining aims at automatically extracting knowledge from unstructured text. Usually the text is organized as a collection of documents, or corpus.

$$\text{TextMining} = \text{NLP} + \text{DataMining}$$

Data Mining aims at automatically extracting knowledge from structured data. (Hand, 2000). Thus, Text Mining is a special case of Data Mining, where input data is text instead of structured data. Normally, text-mining systems create structured representations of the text, which are then analyzed by Data Mining tools. The simplest representation of a text is a vector with the number of occurrences of each word in the text (called a bag-of-words). This representation can be easily created and manipulated, but ignores all the text structure. Text-mining systems may also use Natural Language Processing (NLP) techniques to represent and process text more effectively. NLP is a broad research area that aims at analyzing spoken, handwritten, printed, and electronic text for different purposes, such as speech recognition or translation (Manning and Schütze, 1999). The most popular NLP techniques used by text-mining systems include: tokenization, morphology analysis, part-of-speech tagging, sense disambiguation, parsing, and anaphora resolution.

Tokenization aims at identifying boundaries in the text to fragment the text into basic units called tokens. The first step in a text-mining system is to identify the tokens. The token most commonly used is the word. In most languages, the white-space character can be considered as accurate boundary to fragment the text into words. This problem is more complex in languages without explicitly delimiters, such as Chinese (Wu and Fung, 1994). Morphology analysis aims at grouping the words (tokens) that are variants of a common word, and therefore are normally used with a similar meaning (Spencer, 1991). This involves the study of the structure and formation of words. A common type of inflectional variants results from the tense on verbs. For example, "binding" and "binds" are inflectional variants of "bind." Some other word variants result from prefixing, suffixing, infixing or compounding.

Part-of-speech tagging aims at labeling each word with its semantic role, such as article, noun, verb, adjective, preposition or pronoun (Baker, 1989). This involves the study of the structure and formation of sentences. The tagging is a classification of words according to their semantic role and to their relations to each other in a sentence. Sense disambiguation selects the correct meaning of a word in a given piece of text. For example, “compound” has two different senses in the expressions “compound the ingredients” and “chemical compound.” Normally, the part-of-speech tags are used as a first step in sense disambiguation (Wilks and Stevenson, 1997)

Parsing aims at identifying the syntactic structure of a sentence (Earley, 1970). The syntactic structure of a sequence of words is composed by a set of other syntactic structures related to smaller sequences, except for the part-of-speech tags that are syntactic structures directly linked to words. Normally, the syntactic structure of a sentence is represented by a syntax tree, where leafs represent the words and internal nodes the syntactic structures. Algorithms to identify the complete syntactic structure of a sentence are in general inaccurate and time-consuming, given the combinatorial explosion in long sentences. An alternative is shallow-parsing, which does not attempt to parse complex syntactic structures. Shallow-parsing only splits sentences into phrases, i.e. subsequences of words that represent a grammatical unit, such as noun phrase or verb phrase. Anaphora (or co-reference) resolution aims at determining different sequences of words referring to the same entity. For example, in the sentence “The enzyme has an intense activity, thus, this protein should be used”. The noun phrases “The enzyme” and “this protein” refer to same entity.

Some of the NLP techniques described above can be implemented using algorithms also used in Data Mining. For example, part-of-speech taggers can use Hidden Markov Models (HMMs) to estimate the probability of a sequence of part of speech assignments (Smith, 2004). Not all NLP techniques improve the performance of a given text-mining system. As a result, designers of text-mining systems have to select which NLP techniques would be useful to achieve their ultimate goal.

After creating a structured representation of texts, text-mining systems can use the following approaches for extracting knowledge (Leake, 1996):

Rule-based or Case-based

The Rule-based approach relies on rules inferred from patterns identified from the text by an expert. The rules represent in a structured form the knowledge acquired by experts when performing the same task. The expert analyzes a subpart of the text and identifies common patterns in which the relevant information is expressed. These patterns are then converted to rules to identify the relevant information in the rest of the text. The main bottleneck of this approach is the manual process of creating rules and patterns. Besides being time-consuming, in most cases, this manual process is unable to devise from a subpart of the text the set of rules that encompass all possible cases.

The Case-based approach relies on a predefined set of texts previously annotated by an expert, which is used to learn a model for the rest of the text. Cases contain knowledge in an unprocessed form, and they only describe the output expected by the users for a limited set of examples. The expert analyzes a subpart of the text (training set) and provides the output expected to be returned by the text-mining system for that text. The system uses the training set to create a probabilistic model that will be applied to the rest of the text. The main bottleneck of this approach is the selection and creation of a training set large enough to enable the creation of a model accurate for the rest of the text.

The manual analysis of text requires less expertise in the Case-based approach than in the Rule-based approach. In the Rule-based approach, the expert has to identify how the relevant information is expressed in addition to the expected output. However, Rule-based systems can use this expertise to achieve high precision by selecting the most reliable rules and patterns.

STATE-OF-THE-ART

The main problem in BioLiterature mining is coping with the lack of a standard nomenclature for describing biologic concepts and entities. In BioLiterature, we can often find different terms referring to the same biological concept or entity (synonyms), or the same term meaning different biological concepts or entities (polysyms). Genes, whose name is a common English word, are frequent, which makes it difficult to recognize biological entities in the text.

Recent advances in Text Mining of BioLiterature already achieved acceptable levels of accuracy in identifying gene and protein names in the text. However, the extraction of relationships, such as functional annotations, is still far from being solved. Recent surveys report these advances by presenting text-mining tools that are run in different corpus to perform different tasks (Hirschman, 2002; Blaschke, 2002; Dickman, 2003; Shatkay and Feldman, 2003).

On the other hand, recent challenging evaluations compared the performance of different approaches in solving the same tasks using the same corpus. For example, the 2002 KDD Cup (bio-text task) consisted on identifying which biomedical articles contained relevant experimental results about Drosophila (fruit fly), and the genes (transcripts and proteins) involved (Yeh, 2003). The best submission out of 32 obtained 78% F-measure in the article decision, and 67% F-measure in the gene decision.

A similar challenging evaluation was the 2004 TREC genomics track, which consisted on identifying relevant documents and documents with relevant experimental results about the mouse (Hersh, 2004). The first task was a typical Information Retrieval task. There was given a list of documents and a list of topics. The goal was to identify the relevant documents for each topic. The best submission out of 47 obtained 41% precision. The second subtask comprised the selection of documents with relevant experimental information. The best submission out of 59 obtained 27% F-measure. In addition to document selection, the task also comprised automatic annotations of genes. The best submission out of 36 obtained 56% F-measure.

Another challenging evaluation was BioCreAtIvE (Hirschman, 2005). This evaluation comprised two tasks. The first aimed at identifying genes and proteins in BioLiterature. The best submission out of 40 obtained 83% F-measure. The second task addressed the automatic annotation of human proteins, and involved two subtasks. The first subtask required the identification of the texts that provided the evidence for extracting each annotation. From 21 submissions, the highest precision was 78% and the highest recall was 23%. The second subtask consisted on automatic annotation of proteins. From 18 submissions, the highest precision was 34% and the highest recall was 12%.

AUTOMATIC ANNOTATION

One of the most important applications of text-mining systems is the automatic annotation of genes and proteins. A gene or protein annotation consists of a pair composed by the gene or protein and a description of its biological role. Normally, descriptions use terms from a common ontology. The Gene Ontology (GO-Consortium, 2004) provides a structured controlled vocabulary that can be applied to different species (GO-Consortium, 2004). GO has three different aspects: molecular function, biological process and cellular component. To comprehend a gene or protein activity is also important to know the biological entities that interact with it. Thus, the annotation of a gene or protein also involves identifying interacting chemical substances, drugs, genes and proteins.

Text-mining systems that automatically annotate genes or proteins can be categorized according to: the mining approach taken (Rule-based or Case- based), the NLP techniques applied, and the amount of manual intervention required.

Rule-based Systems

AbXtract was one of the first text-mining systems attempting to characterize the function of genes and proteins based on information automatically extracted from BioLiterature (Andrade and Valencia, 1998). The system assigns relevant keywords to protein families based on a rule comprising the frequency of the keywords in the abstracts related to the family. In addition to using a Rule-based approach, AbXtract relies in only one rule that does not require human intervention. A similar approach is taken by the system proposed by Pérez et al. (2004), which annotates genes with keywords extracted from abstracts based on mappings between different ontologies.

An example of a system based on a large number of rules is BioRAT (Corney, 2004). Given a query, BioRAT finds documents and highlights the most relevant facts in their abstracts or full texts. However, the rules are exclusively derived from patterns inserted by the user. Textpresso is another Rule-based system that finds documents and marks them up with terms from a built-in ontology (Müller, 2004). The system assigns to each entry of the ontology regular expressions that capture how the entry can be expressed in BioLiterature. Textpresso is not so dependent on the user as BioRAT, since many of the regular expressions are automatically generated to account for regular forms of verbs and nouns.

BioIE is a system that takes more advantage of NLP techniques. It extracts biological interactions from BioLiterature and annotates them with GO terms (Kim and Park, 2004). The system uses morphology, sense disambiguation, and rules with syntactic dependencies to identify

GO terms in the text. BioIE uses 1,312 patterns to match interactions in the sentences, so it also requires substantial manual intervention. Koike et al. (2005) propose a similar system that annotates gene, protein and families with GO terms extracted from texts. The system uses morphology, part-of-speech tagging, shallow parsing, and simple anaphora resolution. To extract the relationships, it uses both automatically generated and manually inserted rules.

Case-based Systems

A text-mining system using the Case-based approach was proposed by Palakal et al. (2003). The system extracts relationships between biological objects (e.g. protein, gene, cell cycle). The system uses sense disambiguation, and a probabilistic model to find directional relationships. The model is trained using examples of sentences expressing a relationship.

MeKE is another system that extracts protein functions from BioLiterature using sentence alignment (Chiang and Yu, 2003). MeKE also uses sense disambiguation. The system uses a statistical classifier that identifies common patterns in examples of sentences expressing GO annotations. The classifier uses these patterns to decide if a given sentence expresses a GO annotation.

System	Mining	NLP	Manual
Andrade and Valencia (1998)	Rule-based	nil	nil
Pérez et al. (2004)	Rule-based	nil	nil
Corney et al. (2004)	Rule-based	Low	High
Müller et al. (2004)	Rule-based	Low	Medium
Kim and Park (2004)	Rule-based	Medium	Medium
Koike et al. (2005)	Rule-based	High	Medium

Palakal et al. (2003)	Case-based	Medium	Low
Chiang and Yu (2003)	Case-based	Medium	Low

Table 1: Categorization of some recent text-mining systems designed for automatic annotation of genes and proteins. For each system, the table indicates the mining approach taken, the proportion of NLP techniques used and the proportion of manual intervention needed to generate rules, patterns or training sets.

Discussion

The systems described above show how Text Mining can help curators in the annotation process. Most rely on domain knowledge manually inserted by curators (see Table 1). Domain knowledge improves precision, but it cannot be easily extended to work on other domains and demands an extra effort to keep the knowledge updated as BioLiterature evolves. This approach is time-consuming and makes the systems too specific to be extended to new domains. Thus, an approach to avoid this process is much needed.

GOAnnotator

This section illustrates how text-mining can be integrated in a biological database curation process, by describing GOAnnotator, a tool for assisting the GO annotation of UniProt entries (Rebholz-Schuhmann, 2005). GOAnnotator links the GO terms present in the uncurated annotations with evidence text automatically extracted from the documents linked to UniProt entries.

Figure 1 presents the data flow involved in the processing steps of GOAnnotator and in its interaction with the users and external sources. Initially, the curator provides a UniProt accession number to GOAnnotator. GOAnnotator follows the bibliographic links found in the UniProt database and retrieves the documents. Additional documents are retrieved from the GeneRIF database (Mitchell, 2003). Curators can also provide any other text for mining. GOAnnotator then extracts from the documents GO terms similar to the GO terms present in the uncurated annotations.

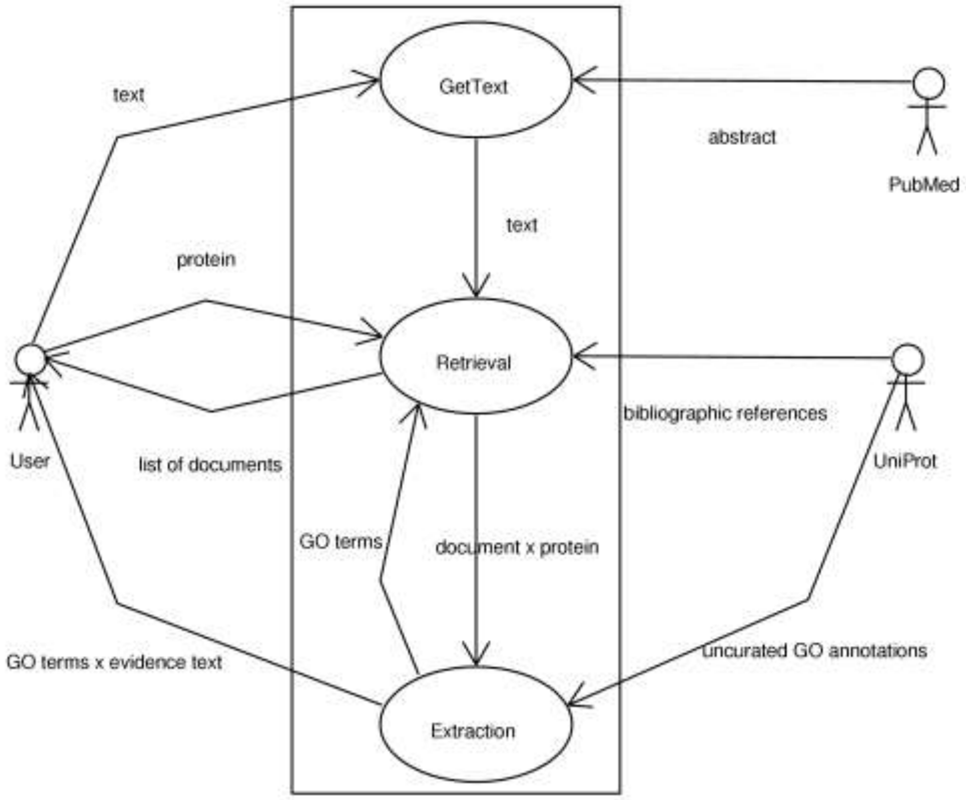
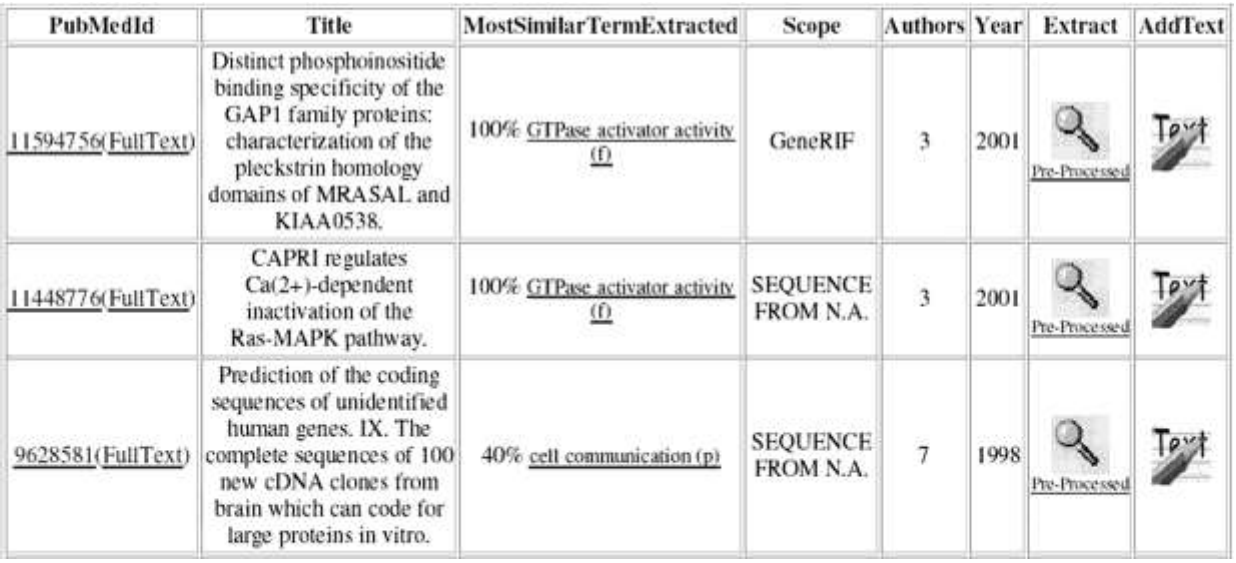
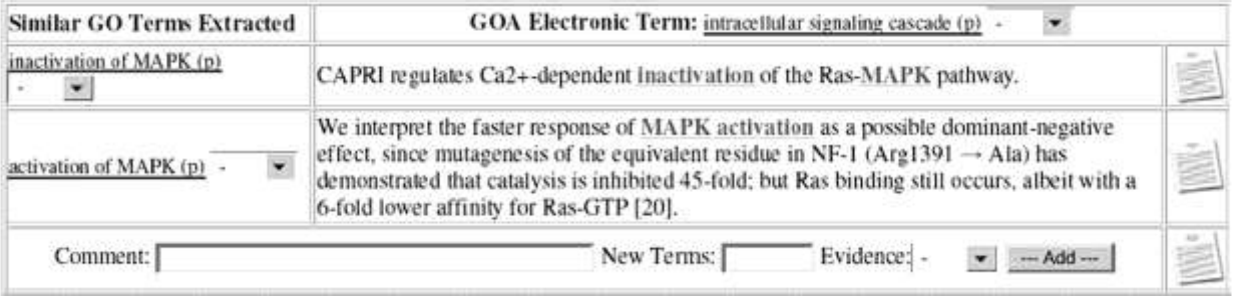


Figure 1: UML use case diagram of GOAnnotator



PubMedId	Title	MostSimilarTermExtracted	Scope	Authors	Year	Extract	AddText
11594756(FullText)	Distinct phosphoinositide binding specificity of the GAP1 family proteins: characterization of the pleckstrin homology domains of MRASAL and KIAA0538.	100% GTPase activator activity (f)	GeneRIF	3	2001	Pre-Processed	Text
11448776(FullText)	CAPRI regulates Ca(2+)-dependent inactivation of the Ras-MAPK pathway.	100% GTPase activator activity (f)	SEQUENCE FROM N.A.	3	2001	Pre-Processed	Text
9628581(FullText)	Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro.	40% cell communication (p)	SEQUENCE FROM N.A.	7	1998	Pre-Processed	Text

Figure 2: A list of documents related to the protein "Ras GTPase-activating protein 4" provided by the GOAnnotator. The list is sorted by the similarity of the most similar term extracted from each document. The curator can invoke the links in the "Extract" column to see the extracted terms together with the evidence text. By default, GOAnnotator uses only the abstracts of scientific documents, but the curator can replace or add text (links in the "AddText" column).



Similar GO Terms Extracted	GOA Electronic Term: intracellular signaling cascade (p) - ▾	
inactivation of MAPK (p) - ▾	CAPRI regulates Ca2+-dependent inactivation of the Ras-MAPK pathway.	
activation of MAPK (p) - ▾	We interpret the faster response of MAPK activation as a possible dominant-negative effect, since mutagenesis of the equivalent residue in NF-1 (Arg1391 → Ala) has demonstrated that catalysis is inhibited 45-fold; but Ras binding still occurs, albeit with a 6-fold lower affinity for Ras-GTP [20].	
Comment: ☐ New Terms: ☐ Evidence: - ▾ --- Add ---		

Figure 3: For each uncurated annotation, GOAnnotator shows the similar GO terms extracted from a sentence of the selected document. If any of the sentences provides correct evidence for the uncurated annotation, or if the evidence supports a GO term similar to that present in the uncurated annotation, the curator can use the "Add" option to store the annotation together with the document reference, the evidence codes and additional comments.

In GOAnnotator the extraction of GO terms is performed by FiGO, a tool that receives text and returns the GO terms detected (Couto, 2005). FiGO is Rule-based, does not use any NLP technique and does not require manual intervention. FiGO assigns a confidence value to each GO term that represents the terms' likelihood of being mentioned in the text. The confidence value is the product of two parameters. The first, called local evidence context (LEC), is used to measure the likelihood that words in the text are part of a given GO term. The second parameter is the inverse of their frequency in GO. GO terms are similar if they are in the same lineage or if they share a common parent in the GO hierarchy. FiGO uses the semantic similarity measure of (Lin, 1998) to compute the degree of similarity between two GO terms.

GOAnnotator ranks the documents based on the extracted GO terms from the text and their similarity to the GO terms present in the uncurated annotations (see Figure 2). Any extracted GO term is an indication for the topic of the document, which is also taken from the UniProt entry.

GOAnnotator displays a table for each uncurated annotation with the GO terms that were extracted from a document and found similar to the GO term present in the uncurated annotation (see Figure 3). The sentences from which the GO terms were extracted are also displayed. Words that have contributed to the extraction of the GO terms are highlighted. GOAnnotator gives the curators the opportunity to manipulate the confidence and similarity thresholds to modify the number of predictions.

FUTURE TRENDS

The performance of text-mining tools that automatically annotate genes or proteins is still not acceptable by curators. Gene or protein annotation is more subjective and requires more expertise than simply finding relevant documents and recognizing biological entities in texts. Moreover, an annotation tool can only perform well when it is using the correct documents and the correct entities. Errors in the retrieval of documents or in the recognition of entities will be the cause of errors in the annotation task.

Existing tools that retrieve relevant documents do not always provide what the curators want. On the contrary, curators spend a large amount of their time finding the right documents. This is probably the main reason why many curators are still not using text-mining tools for gene or protein annotation. Another reason is that quite often the full texts are not electronically available. Curators need additional information that is not usually present in the abstracts, such as the type of experiments applied and the species from which proteins originate. Finally, another reason is that most text-mining tools depend on domain knowledge manually inserted by curators, which is also very time-consuming.

Text-mining tools acquire domain knowledge from the curators in the form of rules or cases. The identification of rules requires more effort to the curators than the evaluation of a limited set of cases. However, a single rule can express knowledge not contained in a large set of cases. Neither source of knowledge subsumes the other: the knowledge represented by a rule is normally not well-represented by any set of cases, and it is difficult to identify a set of rules representing all knowledge expressed by a set of cases.

Couto et al. (2004) proposed an approach to obtain the domain knowledge that does not require human intervention. Instead of obtaining the domain knowledge from curators, they propose acquiring it from publicly available databases that already contain curated data. Text-mining systems could consider these databases as training sets from which rules, patterns or models can be automatically generated. Besides avoiding direct human intervention, these automated training sets are usually much larger than individually generated training sets. Another advantage is that the tools' training data does not become outdated as public databases can be tracked for updates as they evolve.

CONCLUSIONS

Bioinformatics aims at understanding living systems by inferring knowledge from biological information, such as DNA and protein sequences. The role of Text Mining in Bioinformatics is to automatically extract knowledge from BioLiterature. This field is new and has evolved over the last decade, motivated by the opportunity brought by tapping the large amount of information that has been published in BioLiterature with automatic text processing software.

Researchers will tend to use databases to store and find facts, but the evidence substantiating them will continue to be described as unstructured text. As a result, text-mining tools will continue to have an important role in Bioinformatics. Recent advances in Text Mining of BioLiterature are already promising, but many problems remain. In our opinion, the future of text-mining tools for gene or protein annotation will mainly depend on a better use of NLP techniques, and in an efficient integration of existing domain knowledge available in biological databases and ontologies.

REFERENCES

Andrade, M., & Valencia, A. (1998). Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics, 14*(7), 600-607.

Apweiler, R., Bairoch, A., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N., & Yeh, L. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Research, 32*(Database issue), D115-D119.

Baker, C. (1989). *English syntax.* MIT Press.

Blaschke, C., Hirschman, L., & Valencia, A. (2002). Information extraction in molecular biology. *Briefings in BioInformatics, 3*(2), 154-165.

Blaschke, C., Oliveros, J., & Valencia, A. (2004). Mining functional information associated to expression arrays. *Functional and Integrative Genomics, 1*(4), 256-268.

Chiang, J. & Yu, H. (2003). MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics, 19*(11), 1417-1422.

Corney, D., Buxton, B., Langdon, W., & Jones, D. (2004). BioRAT: Extracting biological information from full-length papers. *Bioinformatics, 20*(17), 3206-3213.

Couto, F., Martins, B., & Silva, M. (2004). Classifying biological articles using web resources. *Proceedings of the 2004 ACM Symposium on Applied Computing*, 111-115.

Couto, F., Silva, M., & Coutinho, P. (2005). Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics, 6*(Suppl 1), S21.

Dickman, S. (2003). Tough mining: The challenges of searching the scientific literature. *PLoS Biology, 1*(2), E48.

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of ACM, 13*, 94-102.

GO-Consortium (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Research, 32*(Database issue), D258-D261.

Hand, D., Mannila, H., & Smyth, P. (2000). *Principles of Data Mining.* MIT Press.

Hearst, M. (1999). Untangling text data mining. *Proceedings of the 37th ACL Meeting of the Association for Computational Linguistics*, 3-10.

Hersh, W., Bhuptiraju, R., Ross, L., Johnson, P., Cohen, A., & Kraemer, D. (2004). TREC 2004 genomics track overview. *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*, 14-24.

Hirschman, L., Park, J., Tsujii, J., Wong, L., & Wu, C. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, *18*(12), 1553-1561.

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCrEAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, *6*(Suppl 1), S1.

Kim, J. & Park, J. (2004). BioIE: retargetable information extraction and ontological annotation of biological interactions from literature. *Journal of Bioinformatics and Computational Biology*, *2*(3), 551-568.

Koike, A., Niwa, Y., & Takagi, T. (2005). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, *21*(7), 1227-1236.

Leake, D. (1996). *Case-Based Reasoning: Experiences, Lessons, and Future Directions.* AAAI Press/MIT Press.

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning*, 296-304.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press.

Mitchell, J., Aronson, A., Mork, J., Folk, L., Humphrey, S., & Ward, J. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. Paper presented at the AMIA 2003 Annual Symposia, Washington, DC.

Müller, H., Kenny, E., & Sternberg, P. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLOS Biology, 2*(11), E309.

Palakal, M., Stephens, M., Mukhopadhyay, S., & Raje, R. (2003). Identification of biological relationships from text documents using efficient computational methods. *Journal of Bioinformatics and Computational Biology, 1*(2), 307-342.

Pérez, A., Perez-Iratxeta, C., Bork, P., Thode, G., & Andrade, M. (2004). Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics, 20*(13), 2084-2091.

Rebholz-Schuhmann, D., Kirsch, H., & Couto, F. (2005). Facts from text - is text mining ready to deliver? *PLoS Biology, 3*(2), e65.

Shatkay, H. & Feldman, R. (2003). Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology, 10*(6), 821-855.

Smith, L., Rindflesch, T., & Wilbur, W. (2004). MedPost: a part-of-speech tagger for biomedical text. *Bioinformatics, 20*(14), 2320-2321.

Spencer, A. (1991). *Morphological theory.* Oxford: Blackwell.

Wilks, Y. & Stevenson, M. (1997). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering, 4*(3).

Wu, D. & Fung, P. (1994). Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. *Proceedings of the 4th ACL Conference on Applied Natural Language Processing,* 13-15.

Yeh, A., Hirschman, L., & Morgan, A. (2003). Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. *Bioinformatics*, *19*(1), i331-i339.